Pattern Recognition in Traffic Violations with the Maryland Traffic Violation Dataset

Darrin Daly
School of Computing
National College of Ireland

Francois O Djonga
School of Computing
National College of Ireland

M.D. Harish
School of Computing
National College of Ireland

Weiqi Wang
School of Computing
National College of Ireland

***Abstract*—A Traffic violation is a very common phenomenon in today's world. Traffic violations increase the difficulty for police to manage road safety conditions and increase the risk for severe traffic accidents. With limited resources, dealing with traffic violations is a big challenge for police. Many previous studies have investigated traffic violations to gain more understanding, however, few studies have tried to discover patterns in traffic violations with the aim of helping the police department manage their limited resources. The present study aims to fill this gap by performing data mining techniques with a Montgomery County (Maryland) traffic violation dataset. This dataset contains records of traffic violations with many detailed descriptions which is ideal for the objective of this paper. Specifically, many models (e.g., linear regression, random forest, support vector machine etc.) are implemented to predict the number of severe violations based on date. In addition, ensemble methods are used on these base models. However, the results generated from ensemble methods are no better than the individual models.**

I. Introduction

Traffic violations are an inherent part of modern life. Violations can be positive as a revenue generation source, however, most outcomes are negative. Not only does the steady level of violations pose a heavy burden in economic terms, it also has a more important effect on personal injuries and fatalities of road users. With over 253 million vehicles on the road this social issue is not going anywhere [1]. Although the relative number of violations has remained steady, the number of violations involving fatalities and injuries has decreased slightly in the years between 2010 and 2015. This is a positive trend, however, with yearly national fatalities of over 32,000 and injuries of approximately 500,000 this is still an extremely important social factor within the US.

Taking a more micro level view, for the state of Maryland there were 416 fatalities in 2014 (down from 465 in 2013 and 505 in 2012) yielding a fatality rate of 7.4 per 100,000 population [2]. While this is towards the lower level nationally, it is still a cause for concern. Indeed with the number of registered vehicles in Maryland currently at 2.6 million and continuing to rise, personal injuries and fatalities due to vehicular violations will continue to be a social issue for years to come.

The Montgomery County police department dataset supplemented by a weather dataset taken from the National Climatic Data Center (NCDC) is the basis of this analysis. Numerous papers in the past have focused on classification of traffic accidents. The approach of this paper is different in that it will attempt to predict the number of violations causing injuries or fatalities. Many of the aforementioned papers include large demographic or sensory (smart road) databases. This analysis will be performed using the most basic and readily available police database and weather dataset.

The main goal of this paper is to utilize machine learning to accurately predict future injuries and fatalities. This is achieved by identifying key variables, applying multiple models and then combining them through an ensemble technique. Ultimately, the goal is to accurately predict injuries and fatalities to a degree which will allow for forward planning by emergency services and revenue planning by local councils and indeed count the cost of violations in Montgomery County. For the remainder of this document accidents and violations will be considered interchangeable.

II. Previous Work

Clearly there is a need to reduce traffic accidents and indeed identify the key factors which contribute to them. Expanding this to all traffic violations is an obvious next step. There is a large body of work relating to traffic violations and especially accidents. These papers tend to focus on accident frequency and pedestrian related accidents. For example, Lee and Abdel-Aty [3] completed an analysis regarding vehicle-pedestrian crashes whereby demographic factors, road geometry, traffic levels and environmental conditions were closely related to the severity of pedestrian injuries. While this is related to the topic of this paper, the datasets were more feature rich in terms of demographics and road geometry. The main opportunity from this paper lies in the fact that alcohol was a major contributor to increased injury in both driver and pedestrian situations. A similar study by Zajac and Ivan [4] assessed the severity of pedestrian injuries. While this focuses on pedestrian injuries it does include some relevant contributing features such as alcohol, weather and speed, some of which can be used to inform this papers objective.

Another abstraction of analyses relates to real time traffic control and/or prediction using instrumented freeways [5], [6], [7]. An excellent approach for 'smart' roadways, however, for the majority of situations (e.g. Montgomery County) the only available data is from police records. Thus another method using limited datasets is required in order to accurately predict the number of injuries/fatalities expected in the future.

Other traffic violation papers, such as Durduran [8] and Polat & Durduran [9] focus on the classification of accidents. The former uses a correlation based feature selection method combined with a support vector machine (SVM) and an artificial neural network (ANN) to predict accidents. The latter also uses SVM and ANN and supplements with an ensemble method to achieve a much higher performance. Interestingly it has not been possible to obtain a single paper which attempted to predict the number of future violations, with authors preferring to focus on classification and identifying the main features correlated with accidents.

This paper takes a broader view of personal injuries and fatalities related to all traffic violations including passengers and pedestrians. In addition, this paper attempts to predict future numbers of injuries/fatalities based on simple police gathered data combined with relevant weather data. The main reasons for this approach are threefold; neither demographic nor immediate sensory information is available, the police dataset has a limited feature-set and it appears that this type of general analysis has not previously occurred.

Ensemble methods have been widely adopted and demonstrated great advantages in various data mining competitions (e.g. the Netflix competition and the KDD-cup). The goal of ensemble methods is to build a single model by integrating several different models to produce a greater overall result [10].

To improve the accuracy of this paper, an ensemble type methodology is employed. This ensemble combines four different models chosen due to their dissimilarity from each other and effective fit. It is expected this approach will yield higher returns than any single model approach. A number of combination methods are employed to determine the most appropriate ensemble.

III. Methodology

This is a regression problem which sets out to identify the number of personal injuries and fatalities which will occur within Montgomery County. This is achieved through the judicious use of data mining and statistical methodologies.

A. Knowledge Discovery Process and Workflow

Many researchers have defined knowledge discovery in general, a widely accepted definition is provided by [11] as follows:

> Knowledge discovery in databases (KDD) in the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

In general, there are five stages in a KDD process [12], namely selection, preprocessing, transformation, data mining, and interpretation/evaluation. In addition to these five stages, Chapman [13] proposed an initial stage of gaining business understanding. The tasks of the business understanding stage include determining business objectives, determining data mining goals and so on. The overall workflow of the present project is in line with these procedures. More specifically, some key decisions in the whole process are discussed below.

First, although the present project is an independent academic work, an imaginary client of the project emerges naturally, namely the Montgomery County Police. Consequently, as discussed in the business understanding stage, the objective is to aid the Montgomery County Police to better utilize their limited resources managing traffic violations. More specifically, the data mining goal is to identify some patterns to increase the Montgomery County Police's understanding and ability to predict different aspects of traffic violations.

Although the objective of this project is relatively clear, problems begin to emerge when actually implementing data mining models for the chosen dataset. Initially, the problem the data mining process was defined as a classification problem. Classification techniques such as decision tree and k-means were experimented with at this stage. However, the results were difficult to interpret. In addition, it was realized that even if a classifier was built which could successfully classify different types of violations, there would be no benefit of doing so since violation types are "classified" by the traffic police in reality. Consequently, the data mining objectives must be refined. It is proposed that it should be possible to discover association or correlation patterns between predictors and the target variable (e.g. violation type or severity of the violation). To this end, association rules mining was implemented. Although some rules were generated after experimentation, no particularly valuable rules were identified. Finally, it is proposed that the traffic violation records could be aggregated based on date. The data mining objective is then changed to predict the number of injuries/fatalities based on input variables. In this way, a decision support system could be potentially built to facilitate the management of police resources.

B. Data Collection and Cleansing

As the initial goal was to predict the type of violations based on relevant features, it then mandates that there must be a target variable which contains data regarding types of violations. However, the selected dataset did not have such a target variable. The two columns which contain such information and could potentially be transformed into the target variable are "Description" and "Charge". ETL tools such as OpenRefine, were considered, to process the "Description" column and extract the top ten violation types to be used as the target variable. However, data in the "Description" proved too haphazard to process, since free-text was used to complete the descriptions. An alternative solution presented itself through a website containing traffic violation codes and corresponding titles. Using this data the "charge" for each violation was extracted. Furthermore, it was possible to specify the level of details by extracting only relevant level of codes.

C. Feature Engineering

Due to the initially large number of variables within the dataset, it was necessary to reduce the total variables down to smaller number of variables with higher significance, thereby reducing the noise of the dataset making it easier to identify significance within the dataset.

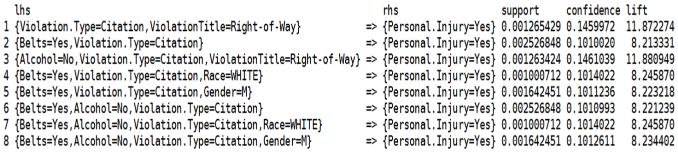
Another problem with the dataset is the class imbalance problem. The class imbalance problem refers to the problem that classes of a variable are represented extremely disproportional in a dataset [14]. Particularly, the number of records that involve injury or fatality are quite small (1.1%). However, as our main data mining objective has changed from classification to prediction of numeric values, this problem then becomes trivial as it mainly affects classification problems [14].

1) Feature Reduction: As the dataset was rich in the numbers of variables, it was necessary to reduce this number to a more manageable level, especially since a relatively large number of algorithms were being used as inputs to an ensemble. Of course the variance must be maintained in order to maintain the integrity of the dataset.

The output of the feature reduction algorithm provided a number of highly significant variables; alcohol, belts, Property.Damage, Commercial.License, HAZMAT, Commercial.Vehicle, Alcohol, Work.Zone, Violation.Type, Weekday, Weekend, Injury.Fatal, year, month, day, weekend & weekday. All other variables were removed based both analysis results and personal knowledge.

2) Association Rules: Frequent patterns are "patterns (e.g., item sets, sub-sequences, or substructures) that appear frequently in a data set" [15]. Frequent pattern mining "searches for recurring relationships in a given data set" [15]. Association rules represent association and correlation between item sets [16]. There more than 20 types of interesting measures (e.g., Kappa, cosine, and Jaccard etc.) for evaluating association rules [17]. However, support, confidence, and lift are the three most commonly adopted measures for assessing interestingness/usefulness and level of certainty among identified rules [16]. Support is the probability of records that contain both event A and B appear in the entire dataset. Confidence is the probability of records that contain event A also contain event B. Finally, lift is confidence divided by the probability of event B.

The initial motivation of experimenting with association rules mining was to discover frequent patterns in our dataset. More specifically, the goal was to discover recurring associations and correlations between target variables and other relevant features. For instance, "Injury" could be regarded as a target variable while "Color", "Belts", "Alcohol", and "HourOfDay" could be used as relevant features. Ideally, the results of association rules could generate results with "Injury" on the right hand side and relevant features on the left hand side. Interesting associations could then be discovered. The process of mining association rules is achieved by using the "arules" package. Initially, the minimum support threshold was set to 0.05 and the minimum confidence threshold was set to 0.8. However, no rules were found by using the combination of these two parameters. Therefore, the parameters have to be modified to identify possible associations. Considering the structure of the dataset, there are approximately 0.1% of records that involve personal injury. Therefore, setting the minimum support threshold (i.e., the probability of records that contain both personal injury and other item sets) too high will not generate any rules. After experimenting several times, a parameter combination of 0.001 support threshold and 0.1 confidence threshold is found to be useful. This configuration generates 16 rules in total. The results are shown in the figure. While these rules do not seem to be highly significant, they do correlate with the output from the feature reduction analysis providing a second layer of support for these important variables to be used as inputs into the various regression models in future sections.



```
  lhs                                                               rhs                   support     confidence lift
1 {Violation.Type=Citation,ViolationTitle=Right-of-Way}            => {Personal.Injury=Yes} 0.001265429 0.1459972 11.872274
2 {Belts=Yes,Violation.Type=Citation}                              => {Personal.Injury=Yes} 0.002526848 0.1010020  8.213331
3 {Alcohol=No,Violation.Type=Citation,ViolationTitle=Right-of-Way} => {Personal.Injury=Yes} 0.001263424 0.1461039 11.880949
4 {Belts=Yes,Violation.Type=Citation,Race=WHITE}                   => {Personal.Injury=Yes} 0.001000712 0.1014022  8.245870
5 {Belts=Yes,Violation.Type=Citation,Gender=M}                     => {Personal.Injury=Yes} 0.001642451 0.1011236  8.223218
6 {Belts=Yes,Alcohol=No,Violation.Type=Citation}                   => {Personal.Injury=Yes} 0.002526848 0.1010993  8.221239
7 {Belts=Yes,Alcohol=No,Violation.Type=Citation,Race=WHITE}        => {Personal.Injury=Yes} 0.001000712 0.1014022  8.245870
8 {Belts=Yes,Alcohol=No,Violation.Type=Citation,Gender=M}          => {Personal.Injury=Yes} 0.001642451 0.1012611  8.234402
```

Fig. 1. Results of associations rules mining

3) Variable Engineering: Two columns (i.e., "DateOfStop" and "TimeOfStop") are mainly used for variable engineering. Date Of Stop was split out into year, month, day, weekday and weekend, while Time of stop was split into various times of the day. Time of stop was not used for prediction purposes during the course of this project, however, it could be an interested next step for the analysis.

4) Dataset Transformation: The available dataset was mainly categorical with the target variable being a binary outcome. In order to translate this dataset into something that would be useful for predicting the rate of injuries/fatalities it was necessary to transform the target variable into a continuous variable. This was achieved by grouping the data by the 'Date of Stop' variable thus creating a continuous dataset of a three year period.

The retained features were based on the features identified as being important from the related work section, such as alcohol and weather, and from the feature selection completed. The full list of retained variables and the relevant variables from the weather dataset can be found in Table I.

TABLE I
AGGREGATED DATASET VARIABLES

Variable 1 - 5	Variable 6 - 10	Variable 10 - 13
Alcohol	Temperature	Violation.Type
PersonalInjury.Fatal	Rainfall	Weekday
Belts	Hazmat	Weekend
Property.Damage	Commercial.Vehicle	
Commercial License	Work.Zone	

IV. DATA MINING TECHNIQUES

The model is created on the training data and predicted using testing data. The split used is 75/25. The implementation is performed using numerous packages within R studio. A series of algorithms are employed to perform regression analysis on the combined traffic violation and weather dataset. The fit of each model is calculated using RMSE and 'Sum-of-Squares'.

RMSE is often used to measure the differences between values (sample and population values) predicted by a model and the values actually observed [18]. In this case, the differences are between the actual test data versus predicted values. RMSE measures the distance between the observation and the fitted line. The square is used to validate negative results. The smaller the value of RMSE the better the fit of the model.

Where possible for every model, each variable is added one at a time in order to gain the optimum r-squared fit while maintaining a strong adjusted r-squared value. The important independent variables are proven through the review of each model. For example, for the linear regression both property.damage and weekday were deemed to be the most important based on their very low P scores.

Outliers can affect negatively regression lines. Prior to the selection of each model, outlier detection was performed. While some of the variables were slightly non-normal in distribution terms, there were no significant outliers within either dataset used in the analyses.

A. Linear Regression

Linear regression is a simple yet powerful and popular supervised machine learning algorithm. A key assumption of a linear regression is that a linear relationship exists between the predictors and the response variable. The application of this algorithm is suitable as both the predictors and the response variable are quantitative in nature. This algorithm approximates the actual function which generates the observed data. As a result the function uses the dataset to 'fit' or 'train' the model. The approach used is the least squares method which is determined by combining all of the error values from the regression model. The terms are squared to eliminate potentially negative values. The smaller the errors the better the linear regression line 'fits' the dataset. The formula is

$$Y = a + bX + e$$

Where X is the explanatory variable and Y is the dependent variable. The associated e is the error for the function. Plotting the residuals versus one of the predictors, in this case alcohol provides Figure 2.

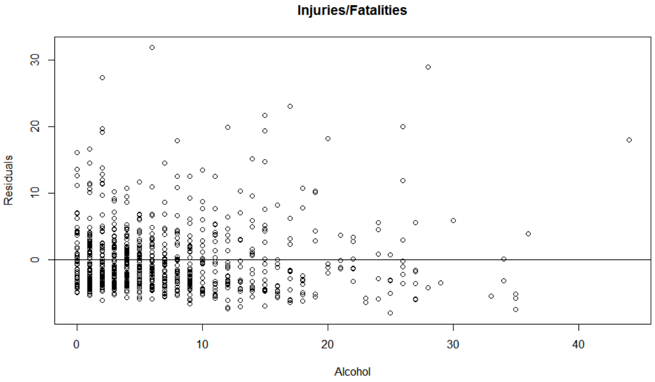
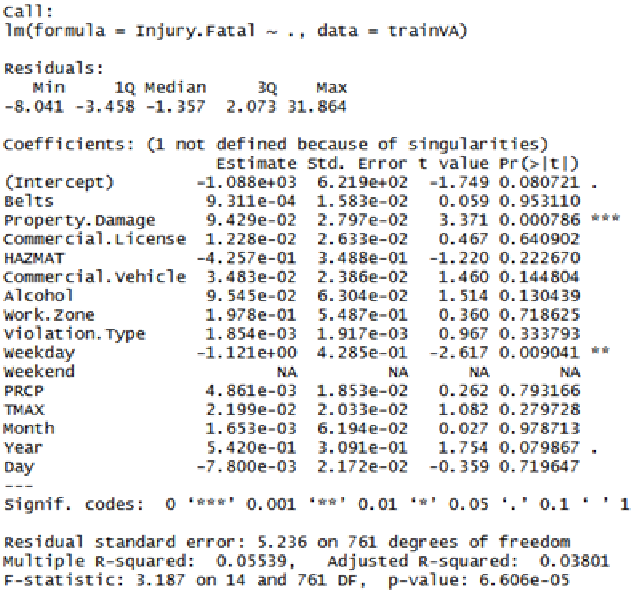


Fig. 2. Residuals of the Linear Regression vs. Alcohol

For a regression it should not be possible to predict the error for a given observation, i.e. the errors should be random. As we can see from Figure 2, there is no real pattern to the residuals, which means that the errors are indeed random. All input variables are plotted and the resultant errors are all random in nature. This shows that there is no obvious bias in the model.

Figure 3 provides a view of the coefficients, residuals and key statistics. It is clear that the most important independent variables for this regression are property.damage and weekday. The model has been optimized based on the r-squared and adjusted r-squared properties. It is important that these two stay close to each other so as not to overfit the model towards the training dataset. If the model is overfit the r-squared will continue to increase while the adjusted r-squared will reduce.



```
Call:
lm(formula = Injury.Fatal ~ ., data = trainVA)

Residuals:
   Min     1Q Median     3Q    Max 
-8.041 -3.458 -1.357  2.073 31.864 

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -1.088e+03  6.219e+02  -1.749 0.080721 .  
Belts               9.311e-04  1.583e-02   0.059 0.953110    
Property.Damage     9.429e-02  2.797e-02   3.371 0.000786 ***
Commercial.License  1.228e-02  2.633e-02   0.467 0.640902    
HAZMAT             -4.257e-01  3.488e-01  -1.220 0.222670    
Commercial.Vehicle  3.483e-02  2.386e-02   1.460 0.144804    
Alcohol             9.545e-02  6.304e-02   1.514 0.130439    
Work.Zone           1.978e-01  5.487e-01   0.360 0.718625    
Violation.Type      1.854e-03  1.917e-03   0.967 0.333793    
Weekday            -1.121e+00  4.285e-01  -2.617 0.009041 ** 
Weekend                    NA         NA      NA       NA    
PRCP                4.861e-03  1.853e-02   0.262 0.793166    
TMAX                2.199e-02  2.033e-02   1.082 0.279728    
Month               1.653e-03  6.194e-02   0.027 0.978713    
Year                5.420e-01  3.091e-01   1.754 0.079867 .  
Day                -7.800e-03  2.172e-02  -0.359 0.719647    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.236 on 761 degrees of freedom
Multiple R-squared:  0.05539,	Adjusted R-squared:  0.03801 
F-statistic: 3.187 on 14 and 761 DF,  p-value: 6.606e-05
```

Fig. 3. Linear Regression Model Detail

Figure 4 shows the regression line when plotted against the most important variable 'property.damage'. This is a positive correlation and the regression line seems to fit quite well. This model must now be applied to the test dataset with the results analyzed using RMSE. For the linear regression the RMSE is 5.4.

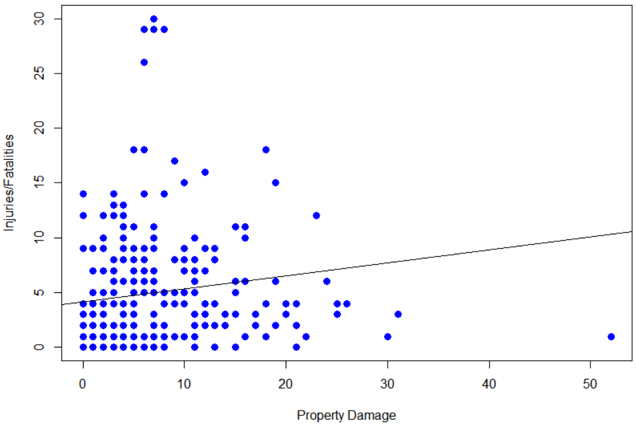


Fig. 4. Linear Regression: Injuries & Fatalities vs. Property Damage

Figure 3 shows the regression line when plotted against the most important variable as pertains to the resulting linear regression model. As we can see this is a positive correlation and the regression line seems t fit quite well. This model must now be applied to the test dataset with the results tested against RMSE.

RMSE is an assessment of the error between the actual and the predicted values. The lower r this value is the better the overall fit of the algorithm. For the linear regression the RMSE is 5.4.

B. Random Forest

Random forest is a data mining technique which can be used for both classification and regression techniques. It defines a classifier for each tree and then ensembles those trees to further improve the overall result for the model. Thus a random forest is a combination of decision trees which are averaged for regression problems, to improve the overall model accuracy [19].

Neither outliers not noise strongly affect the random forest outcome. It is simple, fast and scalable when compared to other ensemble technique such as bagging and boosting. As a result of its bagging mechanism it is resistant to overfitting and it also reduces the generalization error.

The number of trees was set to 100 as no improvement above this level was observed. Mtry was set to 3 as this setting provided the most robust output. The “Date” column was removed from both the datasets since it is not required for performing random forest and more over it was one of the least important variables. Figure 4 shows the predicted versus the actual from the test dataset.

RMSE was calculated as 5.48. RMSE is the difference in between the predicted and the observed variable and it is the “Measure of Fit” of any model. The lower the RMSE the better the model fit. In this case the RMSE value is quite poor. Further analysis in the results section will provide additional assessment of the model.

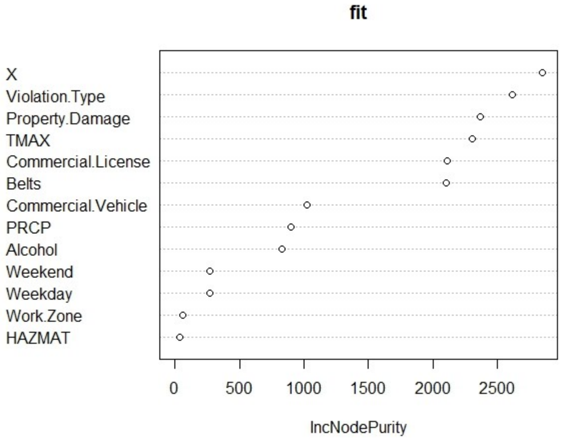


Fig. 5. Random Forest Feature Importance

Figure 5 shows the most important factors relating to the random forest output. The top five features, in order, are Violation type, Property Damage, Temperature, Commercial License which also explains how node impurity has been reduced in each split.

C. Support Vector Machine

Support vector machine (SVM) is a supervised method, essentially a black box, which is a complex but effective data mining technique. It involves statistical and mathematical evaluation and is used for classification and regression purposes for both linear and non-linear data. According to Jiawei Han, Micheline Kamber and Jian Pei [15] the algorithm is explained with the help of non-linear mapping of data into higher dimension. It searches for a boundary or a hyper plane that separates two classes of data within the dimensions. For the regression analysis the model can be expressed in a number of support vectors and can be applied to nonlinear situations using kernel functions. Again as per linear regression the objective is to create a fit with the training data, which minimizes the prediction error.

Similar to a random forest, an SVM model is robust against overfitting and noisy data, therefore producing high accuracy compared to other black box models such as Artificial Neural Network (ANN).

Similar to a random forest, an SVM model is robust against overfitting and noisy data, therefore producing high accuracy compared to other black box models such as Artificial Neural Network (ANN).

The kernel type (rbfdot, polydot, vanilladot) is setup to perform nonlinear mapping on a radial, polynomial and linear basis. The SVM created 681, 680, 775 support vectors respectively. The root mean square error was calculated as per Table II.

TABLE II
SVM RESULTS

Linear	Polynomial	Radial
21.29	22.68	23.73

It can be seen that the linear basis kernel has lowest RMSE result, hence the linear approach produces the best fit of the model. The model was tuned further yielding an RMSE of 5.5 which indicates a far superior fit to the data. The graph below in figure 6 shows the tuned models relationship between cost and epsilon. Here the darker the color, the more accurate the model is.

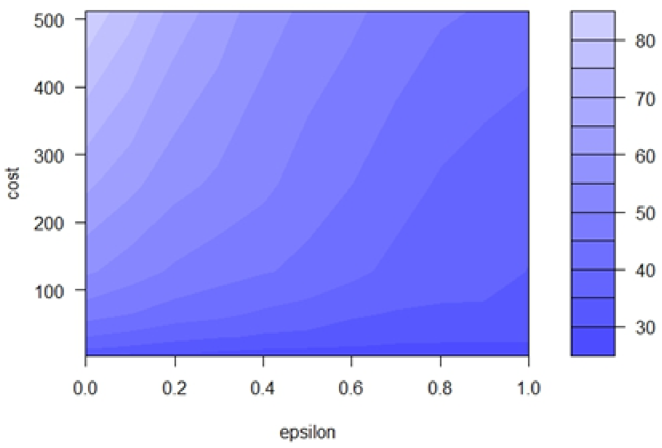


Fig. 6. SVM Tuned Output

Epsilon affects the smoothness and accuracy of the SVM output. Epsilon is inversely proportional to the accuracy of the model. Another parameter in SVM is the cost or C, which determines the fit of the model. If the C value it too high then it is said to experience over-fitting, while too small leads under-fitting, hence a moderate model has a greater effect on the fit. Hence in Figure 6 as cost decreases and epsilon increases, the accuracy of the model improves. It can be said that the as epsilon increases so too does the accuracy of the model.

D. Ordinary Least Square Regression

In statistics, ordinary least squares (OLS) or linear least squares is a process for estimating the unknown parameters in a linear regression model with the purpose of minimizing the differences between the observed responses in a dataset and the responses predicted by the linear approximation of the data [20]. Visually this is seen as the sum of the vertical distances between each data point in the set and the matching point on the regression line - the smaller the differences, the better the model fits the data. The technique can be applied to single or multiple explanatory variables and also categorical explanatory variables which have been appropriately coded. Ultimately, the model seeks to find a set of coefficients for a line/hyper-plane that minimize the sum of the squared errors.

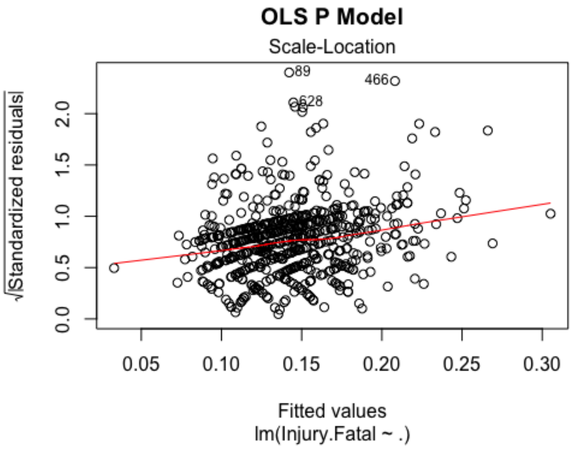
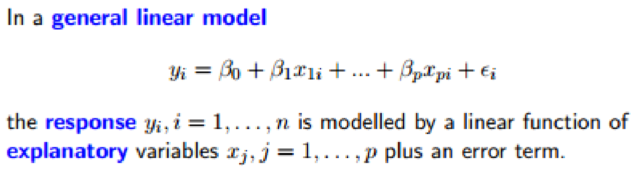


Fig. 7. OLS P Model

The ordinary least square regression also uses mean square error to measure the level of accuracy of the model. RMSE requires that the difference between the test and training should not be too great as otherwise, it is likely that overfitting has occurred. In this case no over fitting has occurred since the training RMSE is lower than the test RMSE. The final RMSE is calculated as 5.13.

E. Generalized Linear Model (GLM)

GLM is a "unified framework for probit models in pursuit of chemical dosage tolerance, contingency tables, OLS regression, and many more" [21]. This generalization of the model offers flexibility to accommodate violations of assumptions for OLS such as non-normal distribution variables. GLM generalizes the regression by using a link function to relate the linear model to the target variable. Thus allowing the magnitude of the variance to be a function of its predicted value, as per Figure 8.



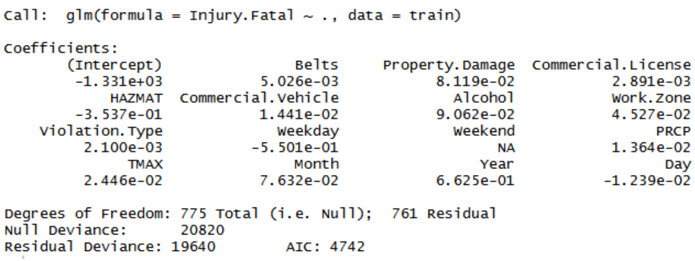
In a **general linear model**

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

the **response** $y_i, i = 1, \ldots, n$ is modelled by a linear function of **explanatory** variables $x_j, j = 1, \ldots, p$ plus an error term.

Fig. 8. GLM Structure [22]

While this is again a linear regression, it is possible that the inclusion of this more flexible version will yield an improved overall ensemble result.



```
Call:  glm(formula = Injury.Fatal ~ ., data = train)

Coefficients:
      (Intercept)               Belts     Property.Damage  Commercial.License
        -1.331e+03          5.026e-03           8.119e-02           2.891e-03
            HAZMAT  Commercial.Vehicle             Alcohol           Work.Zone
        -3.537e-01          1.441e-02           9.062e-02           4.527e-02
    Violation.Type             Weekday             Weekend                PRCP
         2.100e-03         -5.501e-01                  NA           1.364e-02
              TMAX               Month                Year                 Day
         2.446e-02          7.632e-02           6.625e-01          -1.239e-02

Degrees of Freedom: 775 Total (i.e. Null);  761 Residual
Null Deviance:      20820
Residual Deviance: 19640        AIC: 4742
```

Fig. 9. GLM Model

Figure 9 outlines the structure of the GLM model. Each of the inputs has a coefficient. No alterations were required to the datasets to prepare for the application of the GLM model. The RMSE result is 5.37.

F. Neural Network

A Neural Network(NN) is a non-parametric method which mirrors biological neurons to learn from the data [23]. The method uses a series of weights and hidden neurons to detect complex relationships. It performs well in the presence of complicated, noisy or imprecise data. For regression NN is used to predict a multi-dimensional y and share representation, the multi-layer perceptron allows the users to select activation function for the algorithm. In the context of regression analysis, the NN is built to train multi-layer perceptron. The dataset must be normalized for input into the NN to reduce the error level and improve the accuracy of the results. After computing the test then the dataset must de-normalized for built the model.

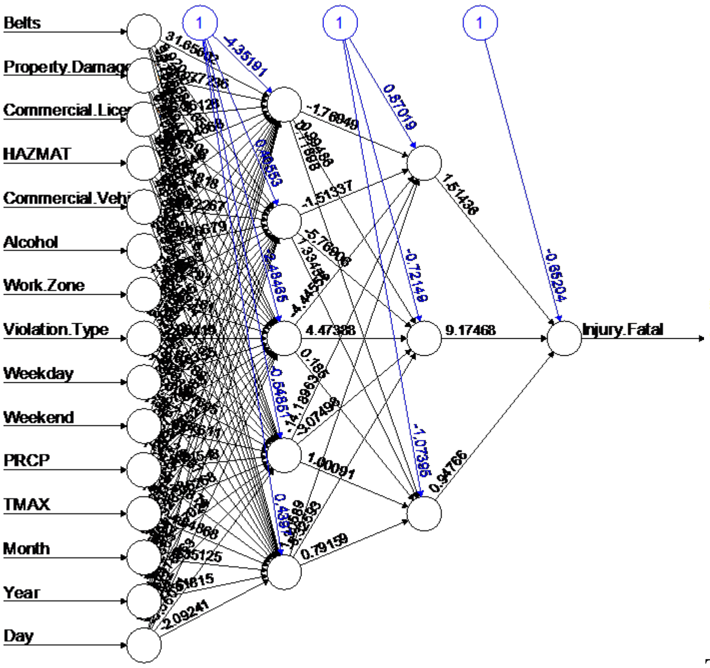


Fig. 10. Neural Network Representation

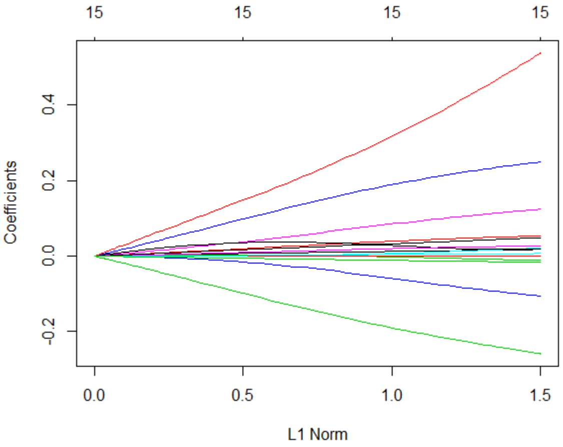
The NN is represented in Figure 10 The number one represents the intercept or constant. The blue figures on each line is the weight of some variable, the weight. The independent variables are the inputs, while the output is a single dependent variable, injury.fatal. Weights are used to determine which connection is more important. These weights are set automatically by backpropagations through a hill climbing algorithm. The RMSE model fit assessment yields a value of 5.93.

G. Regularized Regression

In linear regression models, the ordinary least squares (OLS) estimates are usually used to find the best fit. However, OLS suffers from low prediction accuracy and high complexity for interpretation [24], [25]. In order to deal with the shortcomings of OLS, several regression methods that have been developed in the past decades. Particularly, Ridge regression, the Lasso, and the Elastic Net are the most commonly used ones.

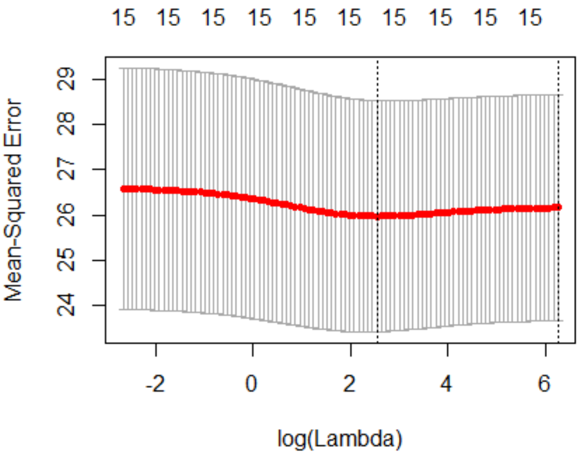
The regularized regression models are implemented by using the “glmnet” package in R. As with other techniques, the target variable is the “Injury.Fatal” column and the predictors. The goal of experiment the model is to find the optimal lambda and alpha for elastic net. When alpha is equal to zero, the model is a ridge model. While alpha is equal to one, the model is a lasso model. The value of alpha for elastic net is greater than zero and smaller than one. In all models, lambda.min is used to make predictions instead of lambda.1se because lambda.1se are much greater than lambda (e.g., in the Ridge model, lambda.min vs. lambda.1se. is 4.06 vs. 815.72).

Fig. 11. Coefficients Plot of Ridge Model



This is a plot of predictor coefficients against lambda. The numbers above the plot are the numbers of nonzero coefficients at different levels of lambda. Each curve in this figure represents a predictor.

Fig. 12. Cross-Validation Curve



This is a cross-validation curve. The two doted lines indicates the value of lambda when the minimum mean cross-validated error.

The table of model coefficients for the three model is shown below. It can be seen that most variables are only partially correlated to the outcome variable.

TABLE III
MODEL COEFFICIENTS

Variables	Ridge	Lasso	Elastic Net
(Intercept)	-277.599	-408.267	-583.2057
Belts	0.006726607	0.009375503	0.01206651
Property.Damage	0.01861966	0.03567608	0.04293509
Commercial.License	0.003039756	0.000	0.000
HAZMAT	-0.01368058	0.000	0.000
Commercial.Vehicle	0.000705462	0.000	0.000
Alcohol	0.03614611	0.0563209	0.08364408
Work.Zone	0.03483951	0.000	0.000
Violation.Type	0.00056482	0.0003474607	0.000389287
Weekday	-0.09314631	0.000	0.1148269
Weekend	0.09314478	0.000	0.0388477
PRCP	0.001380538	0.000	0.000
TMAX	0.009306687	0.01458818	0.02156825
Month	0.01591118	0.000	0.008980206
Year	0.1399793	0.2048086	0.2915966
Day	-0.005022921	0.000	-0.001853043

As suggested by the results, the predictors in general are not highly correlated with the target variable. After comparing RMSE for all implemented alpha values, it is found that RMSE is smallest when the value of alpha is equal to 0.9, which is similar to a lasso model.

H. Ensemble Methods

Essentially, ensemble methods “train multiple learners to solve the same problem” [26]. An ensemble method that uses only one learning algorithm is called a homogeneous ensemble while ensemble methods that integrate multiple different learning algorithms are referred to as heterogeneous ensemble [26]. The main advantage of ensemble methods is that it could generate a strong model by incorporating several weak models [26]. In addition, there are some statistical, computational, and representational reasons for adopting ensemble methods [27]. For example, if the dataset is too big, ensemble methods could split the dataset and train multiple models with the split datasets. Ensemble methods have been widely adopted and demonstrated great advantages in various data mining competitions (e.g., the Netflix competition and the KDD-cup) as well as in real-world applications [28].

With respect to the present project, ensemble methods are adopted for two reasons. First, most individual models implemented do not generate satisfying outcomes. By using ensemble methods, it is expected that the ensembled model have better performance than any individual models as it improves the accuracy and reduces the bias. It would be expected that very different models would have little overlap and therefore would view the dataset from a different perspective. The more models applied in this manner, the more accurate the ensemble should become. Secondly, the use of an ensemble mitigates the uncertainty of algorithm selection, parameter settings and sampling uncertainties. Thirdly, the sample size in our aggregated dataset in relatively small. Using ensemble methods enables the use of re-sampling techniques to deal with this problem.

1) Implementation of the Ensemble Methods: There are various methods which can be used to ensemble. The very basic methods use median, mean, min, max or mode to combine the various prediction outputs from the models to which the dataset is applied. Other more complex methods such as bagging and boosting can also be applied. In this paper all of the basic methods will be applied as well as bagging for comparison purposes.

Bagging, also known as bootstrap aggregation, is an ensemble method whereby an algorithm is developed in which a model is created for multiple sub-sets of data and those models are combined using averaging, for regression. Bagging is mostly used in decision trees, however, not exclusively. As like other models, bagging is robust against overfitting and reduces variance. Bagging is generally used due to its higher accuracy level and ability to handle larger dimensional data.

The only dataset alteration performed was the removal of date column. This model accurately fits the dataset since the Root mean square value was 5.5 and the resampling was performed with cross-validation (10 Folds).

V. RESULTS

This section summarizes & compares the data mining model results and produces a combined output using multiple ensemble techniques. Each of the models is compared using RMSE and ’sum-of-squares’. For both RMSE and ’sum-of-squares’, the lowest value identifies the best model fit for the dataset. Table V shows the full results for all models, while Table VI provides the ensemble results for the various ensemble methods used.

TABLE IV
Model Evaluation: RMSE & Sum-of-Squares

Model	RMSE	Sum-of-Squares
Linear regression (LM)	5.41	7556
Random Forest (RF)	5.57	8014
Lasso	5.34	7372
Ridge	5.33	7340
ElasticNet	5.34	7368
GLM	5.37	7453
Neural Net (NN)	5.93	9069
Ordinary Least Squares (OLS)	5.13	6783
Support Vector Machine (SVM)	5.5	7807

Table V shows that both RMSE and 'Sum-of-Squares' results for all applied models. The OLS model is the most accurately fitted model for the dataset while Lasso, Ridge and ElasticNet are a close second, third and fourth respectively. By far the worst performing model is the Neural Net. One can only assume that the dataset does not suit the Neural Net algorithm. Now looking at Table 4, all of the models from Table 3 have been ensembled using various methods. It is clear that while all 6 methods have similar RMSE outcomes, it is easier to see from the 'Sum-of-Squares' measure that the ever reliable mean value provides the most appropriate fit for the ensembled models. Given that the max ensemble approach shows the least best fit, it would appear that the ensemble techniques are very much on the high side in terms of predictions. Bagging has not provided the expected uplift and remains simply another method which does not provide the best fit.

TABLE V
Ensemble Methods: All Models

Ensemble Method	RMSE	Sum-of-Squares
Min	5.5	7802
Max	5.6	8100
Mean	5.42	7593
Median	5.45	7679
Mode	5.57	8014
Bagging	5.5	7825

Finally, taking the top four models from Table V, combining them and performing ensemble calculations yields Table VII. These 'four' ensemble models perform better than the ensemble with all 9 models present. The full ensemble with all nine models shows the best RMSE result of 5.42 (Mean), whereas the best result for the 'four' model ensemble is 5.21. This indicates that there is extensive overlap between the models and there is no advantage to including all 9 models to obtain the best possible RMSE value. These results, of course, are a relative comparison showing that the 'four' model ensemble solution (Min) is the best fit ensemble for the dataset predictions.

TABLE VI
Ensemble Methods: Top Models Only

Ensemble Method	RMSE	Sum-of-Squares
Min	5.21	7008
Max	5.28	7199
Mean	5.23	7106
Median	5.33	7327
Mode	5.35	7372
Bagging	5.34	7354

Overall, it would appear that a reduced ensemble using the Min method is the best fit for the dataset, however, the RMSE level seems to be quite high indicating that while the Min method is the best fit, it is still not quite as good a fit as one would require to accurately predict the injuries and fatalities. Indeed, the ensemble method performs worse than the individual effort of the OLS model. This would seem to indicate that there is a very high level of overlap in terms of the models explanation of the variance of the dataset. Unfortunately, the combined performance of very different models such as Neural Net and SVM were actually poorer than the linear regression based models. This assertion is supported by the R-squared and adjusted R-squared values of the models whereby less than 60% of the variance is explained in any one model.

Further analysis using cross-validation with a third unseen dataset would provide a more robust accuracy measurement for the final ensemble.

VI. Conclusions

While it has been possible to generate predictions using a number of different algorithms, the accuracy of those algorithms has been left wanting. Firstly, using RMSE it has been possible to asses each of the algorithms separately and then as an ensemble. The output of the ensemble was slightly worse overall when compared to the OLS model. Even taking the OLS model as a standalone model, it still does not account for enough of the variance of the dataset to provide accurate predictions. Secondly, the sum of squares was used to determine which model had the best fit as the difference between the models is much clearer using this measure (even though it is a similar measure to RMSE). Again the OLS model was most prominent as was the final ensemble output using the Min method.

Overall the key objective was to generate predictions with a limited dataset which was supplemented by freely available weather data. This would be the typical scenario in most police department districts in terms of limited available data. The hypothesis was that it would be possible to provide valid and useful predictions for resource management of the local emergency services, and more specifically the police department. In the end the available data is not enough to accurately predict the e level of injuries and fatalities on a daily basis. The dataset must to be supplemented with additional predictor variables which can positively affect the prediction of the target variable. Data such as key demographics (such as age, income, education & job status) as well as key variables such as speed and/or learner driver could be very influential in terms of generating more accurate predictions.

VII. Further Research

The main objective of this paper was to perform a regression analysis to predict the number of vehicle related injuries and fatalities on a daily basis. Further regression research into the time of day would allow a more detailed allocation of emergency services resources. Additionally, a classification of the types of violations which will lead to the injuries or fatalities would provide further insight into the issues that would need to be tackled within Montgomery County in order to reduce the overall level of violations and as a result injuries and fatalities.

A comparative study of the attitudes of other Counties, States or even Countries towards the types and number of citations per capita would provide a benchmark whereby locations could be measured against one another. An interesting offshoot could be a cultural analysis of attitudes towards certain types of traffic violations and a linkage between those violations and costs to the community (both in terms of injuries & fatalities and in monetary terms).

Finally, for this paper it was difficult to procure additional data which could be used to supplement the initial Montgomery dataset. One of the key intentions was to show the level of accuracy achievable with this dataset supplemented by a weather dataset. A comparative analysis with locations (City, County, State, Country) having data such as demographics and/or road sensor data could be used to provide an impetus for local Councils and government agencies to generate and improve their available datasets to benefit the community.

References

[1] Jerry Hirsch, “253 million cars and trucks on U.S. roads; average age is 11.4 years,” Jun. 2014. [Online]. Available: http://www.latimes.com/business/autos/la-fi-hy-ihs-automotive-average-age-car-20140609-story.html

[2] The Insurance Institute for Highway Safety, “Fatality Facts.” [Online]. Available: http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/state-by-state-overview/2012

[3] C. Lee and M. Abdel-Aty, “Comprehensive analysis of vehicleâĂŞpedestrian crashes at intersections in Florida,” *Accident Analysis & Prevention*, vol. 37, no. 4, pp. 775–786, Jul. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457505000564

[4] S. S. Zajac and J. N. Ivan, “Factors influencing injury severity of motor vehicleâĂŞcrossing pedestrian crashes in rural Connecticut,” *Accident Analysis & Prevention*, vol. 35, no. 3, pp. 369–379, May 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457502000131

[5] M. Abdel-Aty and A. Pande, “Identifying crash propensity using specific traffic speed conditions,” *Journal of Safety Research*, vol. 36, no. 1, pp. 97–108, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022437504001197

[6] A. Samimi and B. Hellinga, “Sensitivity of a real-time freeway crash prediction model to calibration optimality,” *European Transport Research Review*, vol. 4, no. 3, pp. 167–174, Jan. 2012. [Online]. Available: http://link.springer.com/article/10.1007/s12544-012-0072-y

[7] C. Lee, B. Hellinga, and F. Saccomanno, “Proactive freeway crash prevention using real-time traffic control,” *Canadian Journal of Civil Engineering*, vol. 30, no. 6, pp. 1034–1041, Dec. 2003. [Online]. Available: http://www.nrcresearchpress.com/doi/abs/10.1139/l03-040

[8] S. S. Durduran, “A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform,” *Expert Systems with Applications*, vol. 37, no. 12, pp. 7729–7736, Dec. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410003684

[9] K. Polat and S. S. Durduran, “Ensemble of classifiers for intelligent recognition of traffic accidents using a geographical information systems platform,” in *1st international symposium on computing in science and Engineering (ISCSE)*, 2010.

[10] L. Rokach, “Ensemble-based classifiers,” *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010. [Online]. Available: http://link.springer.com/article/10.1007/s10462-009-9124-7

[11] U. M. Feyyad, “Data mining and knowledge discovery: making sense out of data,” *IEEE Expert*, vol. 11, no. 5, pp. 20–25, Oct. 1996.

[12] A. I. R. L. Azevedo, “KDD, SEMMA and CRISP-DM: a parallel overview,” *IADS - DM*, 2008. [Online]. Available: http://recipp.ipp.pt/handle/10400.22/136

[13] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0 Step-by-step data mining guide,” Tech. Rep., Aug. 2000. [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf

[14] N. Japkowicz, “The class imbalance problem: Significance and strategies,” in *Proc. of the IntâĂŹl Conf. on Artificial Intelligence*. Citeseer, 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.35.1693&rep=rep1&type=pdf

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, Jun. 2011.

[16] Y. Zhao, *R and Data Mining: Examples and Case Studies*. Academic Press, 2013.

[17] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the Right Interestingness Measure for Association Patterns,” in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’02. New York, NY, USA: ACM, 2002, pp. 32–41. [Online]. Available: http://doi.acm.org/10.1145/775047.775053

[18] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? âĂŞ Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014. [Online]. Available: http://www.geosci-model-dev.net/7/1247/2014/

[19] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1010933404324

[20] L. Leng, T. Zhang, L. Kleinman, and W. Zhu, “Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science,” *Journal of Physics: Conference Series*, vol. 78, no. 1, p. 012084, 2007. [Online]. Available: http://stacks.iop.org/1742-6596/78/i=1/a=012084

[21] J. Dean, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley, May 2014.

[22] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier, Feb. 2011.

[23] F. GÃijnther and S. Fritsch, “neuralnet: Training of neural networks,” *The R Journal*, vol. 2, no. 1, pp. 30–38, 2010. [Online]. Available: https://datajobs.com/data-science-repo/Neural-Net-[Gunther-and-Fritsch].pdf

[24] A. J. Kooij, “Prediction accuracy and stability of regression with optimal scaling transformations,” Ph.D. dissertation, Child & Family Studies and Data Theory (AGP-D), Department of Education and Child Studies, Faculty of Social and Behavioural Sciences, Leiden University, 2007. [Online]. Available: http://openaccess.leidenuniv.nl/handle/1887/12096

[25] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, Apr. 2005. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/abstract

[26] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC Press, Jun. 2012.

[27] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, Aug. 2014.

[28] N. C. Oza and K. Tumer, “Classifier ensembles: Select real-world applications,” *Information Fusion*, vol. 9, no. 1, pp. 4–20, Jan. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253507000620